Studying the Study Section: How Collaborative Decision Making and Videoconferencing Impacts the Grant Peer Review Process

Elizabeth Pier

Joshua Raclaw

Cecilia Ford

Mitchell J. Nathan

University of Wisconsin - Madison

Abstract

Grant peer review is a foundational component of scientific research. In the context of grant review meetings, the review process is a collaborative, socially mediated, locally constructed decision-making task. The current study examines how collaborative discussion impacts reviewers' scores of grant proposals, how different review panels score the same proposals, and how videoconference panels differ from in-person panels. Methodologically, we created and videotaped four "constructed study sections," recruiting biomedical scientists with NIH review experience and an NIH Scientific Review Officer (SRO). These meetings provide a rich medium for investigating the process and outcomes of such authentic collaborative tasks. Implications for research into collaborative decision making as well as for the broad enterprise of federally funded scientific research are discussed.

Objectives

One of the cornerstones of the scientific process is securing funding for one's research. A key mechanism by which funding outcomes are determined is the scientific peer review process. Our focus is on biomedical research funded by the U.S. National Institutes of Health (NIH). NIH spends $30.3 billion on medical research each year, and more than 80% of NIH funding is awarded through competitive grants (NIH, 2015). Advancing our understanding of this review process, including variability between review panels and the efficiency of different meeting formats, has enormous potential to improve scientific research throughout the nation.

NIH's grant review process is a model for federal research foundations, including NSF and IES. It involves panel meetings in which collaborative decision making is an outgrowth of socially mediated cognitive tasks. These tasks include summarizing, evaluating, and critically discussing the perceived scientific merit of applications with other panel members. Investigating how grant review panels function thus allows us not only to better understand processes of collaborative decision making among a group of distributed experts (Brown, Ash, Rutherford, & Gordon, 1993) within a community of practice (Lave & Wenger, 1991), but also to gain insight into the effect of peer review discussions on outcomes for funding scientific research.

Theoretical Framework

A significant body of research (e.g., Cicchetti, 1991; Fogelholm et al., 2012; Langfeldt, 2001; Marsh, Jayasinghe, & Bond, 2008; Obrecht, Tibelius, & D'Aloisio, 2007; Wessely, 1998) has found that inter-reviewer reliability is generally poor in terms of the relative merit of grant applications. Fleurence et al. (2014) and Obrecht et al. (2007) found a general trend of agreement (i.e., score convergence) among reviewers following in-person discussions. Gallo et al. (2013) observed a small improvement in some application scores reviewed through videoconference

compared to in-person meetings, though the meeting format did not significantly impact the reliability or fairness of the review process.

An important framing of the peer review process is the distributed nature of expertise amongst panel members, as some show "ownership" of certain intellectual areas, but no one member can claim it all (Brown et al., 1993). Consequently, co-constructed meanings and review criteria are continually being re-negotiated. Schwartz (1995) showed the advantages of collaborative groups engaged in complex problem solving, whereas Barron (2000) examined some of their variability. Barron noted how groups differentially achieved joint attentional engagement, aligned their goals with one another, and permitted members to contribute to the shared discourse.

Viewing the grant review process through the lens of collaborative decision making via distributed expertise, along with the extant literature, motivates our three research questions:

1. How does the collaborative and distributed discourse during peer review impact reviewers' scores?
2. How consistently do panels of different participants score the same application?
3. In what ways does the videoconference format differ from the in-person format for peer review of grant applications?

Method

As the research team did not have access to actual NIH study sections, we organized four "constructed" study sections comprised of experienced NIH reviewers evaluating recently reviewed applications. Our goal was to emulate the norms and practices of the NIH in all aspects of study design, and our methodological decisions were informed by consultation both with staff

from NIH's Center for Scientific Review (CSR) and with the retired NIH Scientific Review Officer (SRO) who assisted the research team in recruiting grants, reviewers, and chairpersons.

We solicited applications that had been reviewed from 2012 to 2015 by subsections of the Oncology review groups for the National Cancer Institute (NCI). For each application, all research personnel affiliated with an application were assigned pseudonyms and all identifying information was changed. With the SRO's assistance, we recruited 12 reviewers for each of the in-person study sections and eight reviewers for the videoconference study section. For each constructed study section, reviewers were assigned to review six proposals: two as primary reviewer, two as secondary reviewer, and two as tertiary reviewer.

Each study section meeting was organized virtually the same way as an NIH study section. Figures 1 and 2 depict a digitally masked image of screenshots from one of the in-person panels and from the videoconference panel. The SRO begins each panel by convening the meeting, providing opening remarks, and announcing the order of review. Next, the SRO participates throughout the meeting by actively monitoring discussion to ensure that NIH review policy is followed and assists the chair in ensuring there is ample time to discuss all applications. The chair initiates discussion of individual applications by calling on the three assigned reviewers to announce their preliminary scores and verbally summarize their assessments of the application's strengths and weaknesses. The chair then opens the floor for discussion of the grant from both reviewers and non-reviewing panel members. Following discussion, the chair provides a summary of the application's strengths and weaknesses, then calls for the three reviewers to announce their final scores for the application. All panelists then register their final scores using a paper score sheet or, in the case of videoconference meetings, an electronic document. Figure 3 conveys the overall workflow that occurs for a typical study section meeting.

Given our small sample size (42 reviewers nested within four panels—three in-person and one videoconference panel), we take a descriptive approach to these data, as inferential statistics would be severely underpowered. Thus, we utilize descriptive statistics supplemented with qualitative excerpts of discourse from the data to provide an initial, holistic analysis of the processes at play within each of the four constructed study sections.

Data Sources

Data include transcripts from the four panels' verbatim discourse and multiple outcome measures of 20 grant proposals previously reviewed by NIH, including: preliminary impact scores from an application's primary, secondary, and tertiary reviewers submitted prior to the meeting; final impact scores from all panelists submitted during the meeting following discussion; and the time spent discussing each grant. We also compare the final impact scores with those given by the original NIH review panels.

The NIH scoring system uses a reverse nine-point scale, ranging from "Outstanding" (1.0, or a panel-wide impact score of 10) to "Poor" (9.0, or a panel-wide score of 90). Typically, a final impact score of 30 or lower is considered to be a highly impactful project (J. Sipe, personal communication, April 8, 2015).

Results and Discussion

We first examined how peer review affected changes in the scores of the three assigned reviewers. Table 1 lists the average change in scores for each of the grants discussed across all four study sections. Overall, it was more common for the reviewers to worsen their scores for an application after discussion (n=25, 67.57%) than to maintain (n=10, 27.03%) or improve their scores (n=2, 5.41%). There were n=57 (49.14%) instances of an individual reviewer worsening their score, 46 (39.66%) in which they did not change their score, and 13 (11.21%) in which they

improved their score. Thus, individually and in the aggregate, reviewers tended to give less favorable scores following panel discussion. These results provide contrasting evidence to Fleurence and colleagues' (2014) findings that only the weakest applications' scores worsened after discussion, and that discussion itself impacted scores very little. In light of research establishing the benefit of collaboration and group discussion for problem solving (e.g., Cohen, 1994; Schwartz, 1995; Webb & Palinscar, 1996), our findings may indicate a heightened capacity for critically evaluating the merit of grant applications in a collaborative team, as opposed to doing so independently. Future content analyses of panelists' discussions with score changes in mind will be fruitful for exploring this potential explanation.

Exploring our second research question, we found considerable variability in scores across study sections (see Table 2). In Meeting 1, a vast majority of reviewers (71.88%) gave a worse score after discussion, whereas in the other two in-person study sections, reviewers were more evenly split between giving worse scores (38.71% in Meeting 2, 48.48% in Meeting 3) and maintaining their scores (51.61% and 36.36%, respectively). However, in the videoconference meeting, reviewers were more likely to maintain their initial scores (55%) than to worsen (30%) or improve (15%) them (cf. Gallo et al., 2013).

Variability among panels was also evidenced at the individual grant level. However, this was not reflective of a particularly harsh or lenient panel overall. As examples, the Abel and Amsel grants (Table 3, shaded) each received panel impact scores of 27.0 when initially reviewed by NIH. However, each application saw wide variability within our constructed study sections, consistent with findings reported elsewhere (e.g., Barron, 2000; Obrecht et al., 2007).

Our preliminary analysis reveals that one source of variability among panels stems from reviewers' explicitly calibrating scores among one another. For example, in Meeting 2, a panelist

addresses a reviewer: “So it sounds like a lot of weaknesses given that it’s a two [a highly competitive score].” In Meeting 3, a panelist tells the primary reviewer, “Your comments are meaner than your score.” In the videoconference panel, one reviewer remarks, “Yes, so, I respect uh what the rev—the other reviewer said, so I will move my score from two to three.” These examples demonstrate how explicit calibration of scoring norms within a study section directly influences the scoring behaviors of panelists (Langfeldt, 2001), potentially influencing inter-panel reliability for final impact scores.

Research Question 3 compared the videoconference meeting format with the three in-person study section meetings. As Table 3 shows, the average final impact score across all applications was virtually identical for the videoconference meeting compared to Meeting 3, and highly similar to the other in-person meetings, aligning with what Gallo and colleagues (2013) found. Thus, videoconferences do not appear to impair or improve panel outcomes in the aggregate. They may be more efficient, however: The videoconference reviewed eight proposals over two hours and three minutes, while the three in-person panels each reviewed 11 proposals over 2:53, 3:21, and 3:37, respectively. On average (see Table 4), the videoconference panel spent about one minute less *per proposal* than Meeting 1, two-and-a-half minutes less than Meeting 2, and three-and-a-half minutes less than Meeting 3. A correlational analysis between review-time-per-proposal and the average change in panel score showed no discernable pattern, indicating that the time spent on each grant does not strongly predict changes in reviewers’ scores. Future work will investigate the various factors that shape scores in face-to-face versus videoconference formats.

Conclusion

NIH funds nearly 50,000 grant proposals each year. Its study section review process is a model for other federal funding agencies. Constructing study sections designed to emulate the scientific review process is a powerful methodological approach to understanding how scientific research is funded and offers valuable insights into the important area of complex group decision making. These preliminary findings already contribute to scientific understanding of the review process and to policy recommendations for future review panels. We found that panelists are continually renegotiating and recalibrating the meaning of their numerical scores in terms of the norms of the other panel members and the quality of the reviewed proposals. A clear recommendation for training panelists on these scoring procedures, and the development of community-wide norms, would likely improve the consistency of scores across panels. We also found that discussion time is reduced among videoconference panelists, but they perform comparably to in-person panels on average, providing positive support for technology measures to reduce costs, improve efficiency, and increase panelist participation.

The vast undertaking of recruiting and convening expert panels of biomedical researchers puts constraints on the number of panelists and applications for this initial study, limiting our sample size and thus power for conducting inferential statistics. Videoconferencing could increase the number of constructed study sections we can investigate in the future. In future analyses, we will examine the role of chairs in moderating discussion and the factors implicated in score changes (e.g., number of positive and negative comments). Ultimately, in keeping with the AERA 2016 theme promoting democracy through public scholarship, we believe research of this type can increase the public perception of scientific research activities in the U.S. and the role scientific research can play for benefiting public policy.

References

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Sciences*, *9*(4), 403–436.

Brown, A. L., Ash, D., Rutherford, M., & Gordon, A. (1993). Distributed expertise in the classroom. *Distributed Cognitions: Psychological and Educational Considerations*, 188–228.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Bran Sciences*, *14,* 119–135.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, *64*, 1–35.

Fleurence, R. L., Forsythe L. P., Lauer, M., Rotter, J., Ioannidis, J. P., Beal, A., Frank, L., Selby, J. V. (2014). Engaging patients and stakeholders in research proposal review: The patient-centered outcomes research institute. *Annals of Internal Medicine*, *161*(2), 122–130.

Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A., & Väänänen, K. (2012). Panel discussion does not improve reliability of peer review for medical research grant proposals. *Journal of Clinical Epidemiology*, *65*(1), 47–52. doi:10.1016/j.jclinepi.2011.05.001

Gallo, S. A., Carpenter, A. S., & Glisson, S. R. (2013). Teleconference versus face-to-face scientific peer review of grant application: Effects on review outcomes. *PLOS ONE, 8*(8), 1–9.

Langfeldt, L. (2001). The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science*, *31*(6), 820–841. doi:10.1177/030631201031006002

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press. doi:10.2307/2804509

Marsh H. W., Jayasinghe, U. W., Bond N. W. (2008). Improving the peer review process for grant applications: Reliability, validity, bias and generalizability. *American Psychologist*, *63*(3), 160–168.

Obrecht, M., Tibelius, K., & D'Aloisio, G. (2007). Examining the value added by committee discussion in the review of applications for research awards. *Research Evaluation*, *16*(2), 70–91. doi:10.3152/095820207X223785

National Institutes of Health (NIH). (2015, January 29). *NIH Budget.* Retrieved from http://www.nih.gov/about/budget.htm

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *The Journal of the Learning Sciences*, *4*(3), 321–354.

Webb, N., & Palinscar, A. S. (1996). Group processes in the classroom. In R. Calfee & C. Berliner (Eds.), *Handbook of Educational Psychology* (pp. 841–873). New York: Prentice Hall.

Wessely, S. (1998). Peer review of grant applications: What do we know? *Lancet*, *352*, 301–305.

Figures

Figure 1. Digitally masked screen shot depicting the layout of the panel members during Meeting 1.

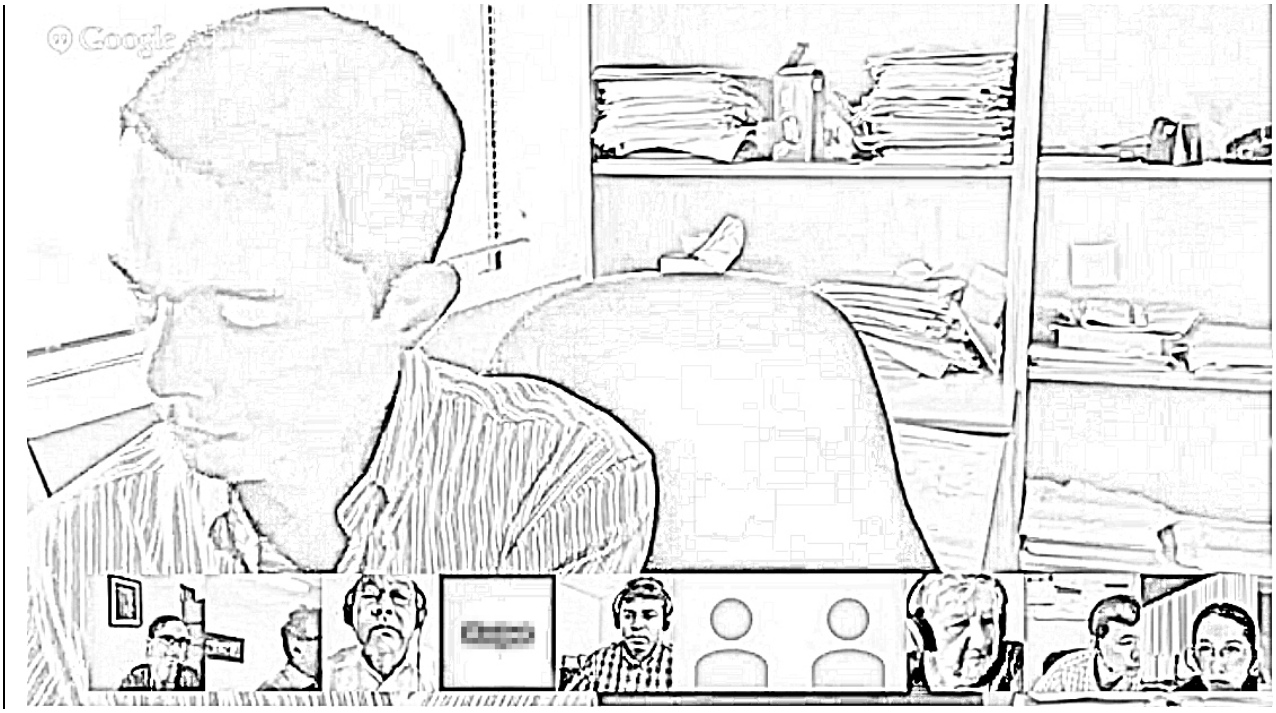

Figure 2. Digitally masked screenshot depicting the panelists' view of the videoconference meeting. The name of the current speaker, featured in the main window, is displayed in the smaller window along the bottom row (fourth from the left) where he would appear when not speaking, but it has been blurred out here for privacy purposes.

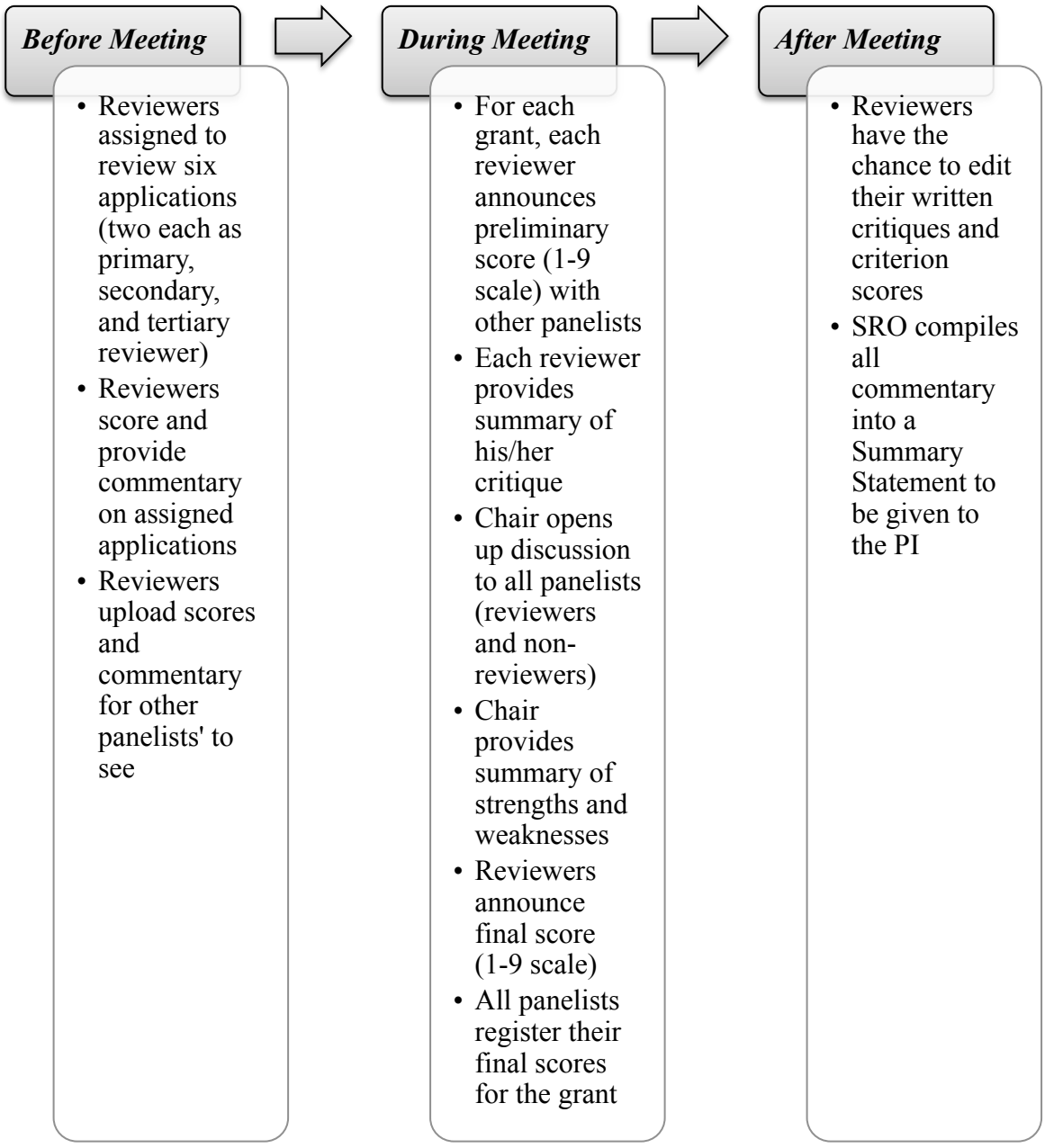

Figure 3. Typical workflow involved in a NIH study section meeting.

Tables

Table 1

Average Changes in Individual Reviewers' Scores Before and After Discussion of Each Grant

Grant	Meeting 1	Meeting 2	Meeting 3	Videoconference
Abel	0	0	–2.333	
Adamsson	–0.667			
Albert	–0.667			–0.333
Amsel	–2	–0.667	+0.333	
Bretz			–0.667	
Edwards		–1.333		
Ferrera			–0.667	
Foster	–1.333	–0.667	0	–1*
Henry	–2.333	–0.333	–0.333	–1
Holzmann				0
Lopez	0	0	+0.333	
McMillan			–1	
Molloy	–2	0		
Phillips	–1	–0.5		
Rice		–1.0	–0.333	
Stavros		–0.667		0
Washington	–0.333	0		+0.333*
Williams	–0.333		–0.333	0*
Wu				+0.667*
Zhang			0	

Note. Grants are labeled by the last name of the PI's pseudonum. N/A indicates a grant that was not discussed at a given meeting due to triaging. *Indicates that these are not exclusively within-subject comparisons, as mail-in reviews were used here due to a reviewer being unable to participate in the study section. Thus, these are excluded from consideration.

Table 2

Count of Changes in Individual Reviewers' Scores Before and After Discussion

Change	Meeting 1	Meeting 2	Meeting 3	Videoconference	Total
Improved *(lower score)*	2 (6.25%)	3 (9.68%)	5 (15.15%)	3 (15.00%)	13 (11.21%)
No change	7 (21.88%)	16 (51.61%)	12 (36.36%)	11 (55.00%)	46 (39.66%)
Worsened *(higher score)*	23 (71.88%)	12 (38.71%)	16 (48.48%)	6 (30.00%)	57 (49.14%)

Table 3

Final Impact Scores

Grant	Meeting 1	Meeting 2	Meeting 3	Video-conference	Average	NIH
Abel	20.0	29.1	50.0		33.0	27.0
Adamsson	30.0				30.0	23.0
Albert	35.0			38.6	36.8	39.0
Amsel	50.0	25.5	20.9		32.1	27.0
Bretz			39.2		39.2	20.0
Edwards		37.3			37.3	40.0
Ferrera			33.3		33.3	36.0
Foster	42.0	38.2	29.2	45.0	38.6	23.0
Henry	52.0	35.5	35.0	32.5	38.8	14.0
Holzmann				27.5	27.5	17.0
Lopez	30.0	21.8	16.7		22.8	39.0
McMillan			30.8		30.8	25.0
Molloy	50.0	30.0			40.0	28.0
Phillips	31.1	30.8			31.0	23.0
Rice		39.1	31.7		35.4	ND
Stavros		32.7		33.8	33.3	20.0
Washington	39.0	35.0		26.3	33.4	31.0
Williams	42.0		30.8	38.8	33.9	28.0
Wu				20.0	20.0	44.0
Zhang			29.2		29.2	38.0
Average	38.3	32.3	31.5	31.6	32.8	28.5

Note. Abel and Amsel grants (shaded) are examples of applications with highly variable final impact scores across constructed study sections.

Table 4

Total Time in Minutes and Seconds Spent on Each Application at Each Meeting

Grant	Meeting 1	Meeting 2	Meeting 3	Videoconference	Average
Abel	17:39	16:43	14:33		16:18
Adamsson	15:0				15:00
Albert	13:18			13:24	13:21
Amsel	13:08	12:14	20:16		15:13
Bretz			15:20		15:20
Edwards		11:01			11:01
Ferrera			20:22		20:22
Foster	14:46	14:58	15:00	09:29	13:33
Henry	15:41	17:27	14:01	20:17	16:52
Holzmann				15:50	15:50
Lopez	18:58	18:24	17:10		18:11
McMillan			25:47		25:47
Molloy	13:22	09:12			11:17
Phillips	13:37	13:17			13:27
Rice		14:02	13:05		13:33
Stavros		19:52		10:20	15:06
Washington	14:27	31:58		13:30	19:58
Williams	13:33		17:07	15:10	15:17
Wu				12:46	12:46
Zhang			17:41		17:41
Average	14:52	16:17	17:18	13:51	15:48